Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 167-170 recibido 26-03-2019 revisado 03-05-2019 aceptado 15-05-2019

Estudio sobre documentos reutilizables como recursos lingüísticos en el marco del desarrollo del Plan de Impulso de las Tecnologías del Lenguaje

Report on reusable documents as language resources in Spain, under the Government Plan for Language Technologies

Antonio Moreno[1], Doroteo Torre[1], Ana Valverde[1,3], Leonardo Campillos[1,2]
[1] Universidad Autónoma de Madrid
[2] LIMSI-CNRS (Orsay, Francia)
[3]Real Academia Nacional de Medicina
{antonio.msandoval, doroteo.torre, leonardo.campillos}@uam.es, avalverde@ranm.es

Resumen: Este estudio ha sido realizado dentro del ámbito del Plan de impulso de las Tecnologías del Lenguaje (Plan TL) con financiación de la Secretaría de Estado para el Avance Digital y Red.es. Los objetivos centrales son realizar un censado de recursos de las diferentes administraciones públicas que puedan ser convertidos en RL, así como proponer un plan de acción para abordar su conversión en RL. Se ha elaborado una metodología específica para el censado y evaluación de la madurez de los datos. Se han generado dos listados, uno preliminar compuesto por 101 recursos, del que se han seleccionado 24 para su análisis detallado y evaluación. El informe también incluye un repaso de estudios similares en otros países. Concluye con unas recomendaciones genéricas, así como estrategias concretas para los recursos seleccionados. El informe final y los listados están disponibles públicamente en Red.es y la página del Plan TL.

Palabras clave: recursos lingüísticos, Plan TL, administraciones públicas, datos abiertos, reutilización de información del sector público, RISP.

Abstract: This report was carried out within the Spanish administration-driven initiative Language Technologies Plan (Plan TL), funded by Secretaría de Estado para el Avance Digital and Red.es. The main goals are collecting from Spanish public administrations a listing of provided resources and open data that can be transformed to language resources, as well as proposing an action plan to process and distribute them. We designed a specific methodology for listing and evaluating the degree of maturity of the considered data. We created two listings: a preliminary collection of 101 resources, and 24 resources and data repositories selected from the first list for a detailed analysis and evaluation. This report also features a comparative analysis of similar initiatives and studies conducted abroad. We conclude with generic recommendations and detailed strategies for the selected resources. The report and listings are publicly available at Red.es and the Plan TL. website.
Keywords: language resources, Plan TL, Spanish public administrations, Open Data, Public Sector Information, PSI.

1 Contexto del estudio

El Plan TL tiene como objetivo general desarrollar la industria del Procesamiento del Lenguaje Natural (PLN), la Traducción Automática (TA) y los sistemas conversacionales (SSCC) en España, y especialmente, en lengua española y lenguas cooficiales, aprovechando, además, estas novedosas tecnologías para mejorar el servicio público.

En el marco de este Plan, ya se han realizado otros informes (Bel y Rigau 2015; Soroa et al. 2017) sobre el estado de las Tecnologías del

ISSN 1135-5948. DOI 10.26342/2019-63-22

Lenguaje (TL) en España. El objetivo del presente estudio es localizar conjuntos de datos o documentos de distintos organismos públicos que resulten de interés desde el punto de vista de las TL y que sean, además, factibles para su conversión en RL.

2 Metodología empleada

Para la elaboración y evaluación del censo de recursos, se generó una ficha técnica para la recogida de información, compuesta por diferentes campos (Tabla 1)

1. **Identificación del recurso:** nombre, url, clasificación, lenguas, licencia, etc.
2. **Persona de contacto u organización responsable**
3. **Creación del recurso:** proveedor, proyecto financiador.
4. **Descripción del recurso:** variedad de lengua, niveles de anotación, estándares, tamaño, unidad, formato, dominio, etc.
5. **Otros recursos relacionados.**
6. **Grado de madurez de datos conforme al modelo de la tabla 2.**
7. **Posibles aplicaciones del futuro recurso lingüístico.**

Tabla 1: Estructura y principales contenidos de la ficha técnica

Adicionalmente, se elaboró una ficha específica para la evaluación de la madurez (Tabla 2), teniendo en cuenta factores técnicos requeridos por los RL usados en PLN y aspectos legales para su (re)utilización.

3 Censado de recursos

3.1 Listado de documentos analizados por dominio

Los 101 recursos recogidos en la primera búsqueda se clasificaron en cuatro áreas temáticas que son de interés prioritario para el Plan TL:

- **Sanidad: 41** (textos, documentos multimedia, entidades nombradas, memorias bilingües, o terminología, de ámbito español o latinoamericano).
- **Justicia: 5** (textos y documentos multimedia de ámbito español).
- **Inteligencia competitiva: 7** (entidades nombradas, terminología y corpus paralelos de ámbito español o latinoamericano).
- **Cultura, Turismo y otros: 35** (textos, documentos multimedia, entidades nombradas y corpus paralelos de ámbito español).

Puntos que considerar para evaluar el grado de madurez	Valores: 0, 1, 2	Observaciones
Aspectos técnicos:		
1. Digitalización (conversión a formato procesable).		
2. Transcripción (ortográfica, fonológica, …).		
3. Alineación vídeo/sonido y texto.		
4. Procesamiento estandarizado y homogéneo de codificación de caracteres (UTF-8 o ISO-8859-1).		
5. Anotación morfológica y/o sintáctica.		
6. Anotación de entidades nombradas.		
7. Otros tipos de anotación (semántica, pragmática, palabras clave, …).		
8. Revisión de aspectos formales (ortografía, formato de etiquetado, …).		
9. Revisión de contenido (incoherencias, redundancia de datos, …).		
10. Anotación conforme a estándares PLN.		
11. Presencia de metadatos.		
Aspectos legales:		
12. Necesidad de anonimización de datos personales.		
13. Necesidad de solicitud de permiso de uso.		
TOTAL		

Tabla 2: Plantilla para la evaluación de la madurez como RL de un recurso

De este primer censo, se seleccionaron 24 recursos para su análisis exhaustivo, según los criterios de interés (calidad, cantidad y disponibilidad de los datos), plurilingüismo[1], estado de la propiedad intelectual, variedad temática, grado de madurez y tipología del RL (ver Tabla 3).

4 Conclusiones preliminares sobre la madurez

Como era de esperar, **la mayoría de los conjuntos de datos analizados** en este informe se quedan en los estadios de **madurez baja** o **media**. Esto es comprensible y esperable, dado

[1] Se han considerado únicamente recursos en las cuatro lenguas cooficiales del Estado.

que los requisitos para ser considerado un recurso *maduro* son muy estrictos: solo los ya procesados y en formatos directamente usables por los investigadores de PLN (por ejemplo, XML o TMX) pueden ser considerados propiamente RL.

Listado de Recursos Analizados
1. Patentes, modelos de utilidad e informes técnicos digitalizados de la Oficina Española de Patentes y Marcas (OEPM).
2. Patentes multilingües digitalizadas en PATSTAT de European Patent Office (EPO).
3. Diccionarios terminológicos del Centro de Terminología (TERMCAT).
4. Padrón: Relación de municipios del Instituto Nacional de Estadística.
5. Topónimos del Instituto Geográfico Nacional (IGN).
6. Grabaciones de vídeo de RTVE a la carta.
7. Grabaciones de audio y vídeo del Archivo Audiovisual del Congreso de los Diputados de España.
8. Índices de clasificación de los catálogos de la BNE.
9. Publicaciones periódicas digitalizadas de la Hemeroteca Digital.
10. Documentos digitalizados de la Biblioteca Digital Hispánica.
11. Publicaciones en repositorio SciELO (Scientific Electronic Library Online).
12. Publicaciones y vídeos del Instituto de Salud Carlos III (ISCIII).
13. Banco de datos de enfermedades raras y medicamentos huérfanos de OrphaData.
14. Guías de práctica clínica (GPC) del portal Guía Salud.
15. Vídeos del portal web de TV del Gobierno Vasco relacionados con el tema de Salud.
16. Publicaciones de la Agencia Española de Medicamentos y Productos Sanitarios (AEMPS).
17. Nomenclátor de prescripción del Centro de Información de Medicamentos (CIMA).
18. Textos de Jurisprudencia del CENDOJ.
19. Textos del Boletín Oficial del Estado (BOE) Diario.
20. Textos de códigos electrónicos del Boletín Oficial del Estado (BOE).
21. Textos sobre Legislación del Boletín Oficial del Estado (BOE).
22. Memorias de traducción que contienen las publicaciones en el Boletín Oficial del Estado realizadas en euskera del Instituto Vasco de Administración Pública (IVAP).
23. Memorias públicas de traducción de la Diputación Foral de Gipuzkoa.
24. Grabaciones de Vistas Judiciales del Consejo General del Poder Judicial.

Tabla 3: Recursos analizados en el estudio

5 Estudios e iniciativas similares en otros países

El estudio contempla iniciativas de creación de RL en países hispanoamericanos y en Europa, con algunas incursiones en EEUU y Canadá. Otros países destacables (China, Japón o India) quedaron fuera de este análisis.

Un plan de acción similar para la reutilización de datos públicos o de contenidos de páginas webs en administraciones públicas para desarrollo de RL es el portal de la European Language Resource Coordination,[2] que proporciona una lista de los recursos existentes en Europa de carácter abierto.

España se encuentra entre los países de cabeza en cuanto a RL en formato libre y disponible para TL. A nivel europeo, Francia y Reino Unido son los referentes, si bien la Comisión Europea ha advertido del retraso con respecto al inglés y a las compañías americanas, que han *capitalizado* el uso de los *Big Data* lingüísticos, permitiendo que las grandes multinacionales tecnológicas como Google, Microsoft, Amazon o IBM ofrezcan servicios lingüísticos en muy diferentes dominios y lenguas.

Pese a la apertura y disponibilidad aparente de recursos, es importante destacar que **la comunidad investigadora en PLN en España necesita RL de calidad y en cantidad suficiente para desarrollar aplicaciones competitivas en el mercado internacional**, con la amenaza potencial de que otras empresas de fuera de nuestras fronteras ocupen el espacio del PLN en las lenguas del Estado.

De ahí la necesidad de que la conversión a RL siga estándares técnicos internacionales para garantizar la interoperabilidad entre datos y procesadores (por ejemplo, los requisitos definidos por CLARIN).

6 Recomendaciones

Otro de los objetivos del estudio es describir un plan de acción a corto y medio plazo, con el fin de establecer prioridades.

6.1 Recomendaciones genéricas

Se aplican a todos los tipos de RL:

1. Garantizar la disponibilidad y el acceso universal a los datos abiertos para RL en todas las lenguas del Estado a través de un portal común y único.
2. Mejorar la visibilidad de los conjuntos de datos en cuanto a su disponibilidad y madurez.

[2] www.lr-coordination.eu

3. Facilitar la descarga masiva de grandes ficheros en formatos apropiados (texto plano, XML, CSV, JSON, RDF).
4. Impulsar la conversión progresiva de los millones de páginas digitalizadas en PDF o imagen en texto plano: los documentos en PDF, EPUB o las imágenes digitalizadas a formato TXT, CSV o XML.
5. Proporcionar la transcripción de ficheros multimedia. Las cadenas públicas de radio y televisión y las grabaciones de vistas orales de los juicios son una valiosa fuente de datos para distintos ámbitos de procesamiento de habla.
6. Estimular la adopción de licencias de libre uso y acceso a los datos. El empleo de licencias de tipo Creative Commons resulta muy práctico y, en términos generales, agiliza la obtención de los RL.
7. Estimular la reutilización de datos organizando competiciones tecnológicas.
8. Facilitar el acceso a capacidad de cómputo y almacenamiento, ya que la evolución de la tecnología en los ámbitos del PLN se apoya, cada vez más, en estos aspectos.

6.2 Estrategias concretas para recursos seleccionados

El informe destaca una serie de conjuntos como líneas prioritarias en el Plan de Acción:

1. Los **recursos digitales de la Biblioteca Nacional de España (BNE)**: tanto la Hemeroteca Digital como la Biblioteca Digital Hispánica poseen conjuntos de datos de enorme interés lingüístico y cultural, de gran variedad temática, temporal y geográfica. Además, están ya digitalizados en PDF, son de acceso libre, no tienen impedimentos legales de copyright, y su uso no requiere la solicitud expresa de un permiso escrito.
2. **Archivo de RTVE** y **RTVE a la carta**: datos (en formato vídeo y audio) de gran interés lingüístico, cultural y tecnológico, con gran variedad temática. Contienen cerca de 100.000 horas de televisión de alta calidad, con más de 10.000 informativos.
3. **Patentes del ámbito iberoamericano** o registradas en la **OEPM** y **PATSTAT**. Si bien el grado de madurez de estos recursos (patentes, modelos de utilidad e informes técnicos) puede ser mejorado para las tareas de PLN, el potencial y el interés es enorme.
4. **Recursos de dominio médico**: destacan algunos recursos de un grado alto de madurez (los bancos de datos de OrphaData y el Nomenclátor de prescripción del CIMA), con un modelo de licencia, distribución y riqueza de anotaciones. Habría un proceso de conversión más costoso para las publicaciones científicas de ISCIII o la AEMPS (junto con sus alertas médicas y farmacológicas), o las guías de práctica clínica del portal GuíaSalud.

7 Conclusiones

Los patrimonios documental, jurídico, cultural y lingüístico del país están muy exhaustivamente representados en diferentes conjuntos de datos abiertos y la comunidad I+D está esperando con fundadas expectativas que se conviertan en RL. Su disponibilidad debería permitir a la industria española de TL poder competir con las empresas internacionales en un Mercado Único Digital basado en el tratamiento multilingüe. De esta manera, España se convertiría en un referente en TL a nivel europeo junto a Francia y Reino Unido, si se potencia el desarrollo y las infraestructuras de RL.

Agradecimientos

Este informe ha sido financiado por la Secretaría de Estado para el Avance Digital (SEAD) y Red.es. Agradecemos las observaciones realizadas por David Pérez, Sonia Castro, Pilar Polo, Inés Rodríguez y Doaa Samy.

Bibliografía

Bel, N. y G. Rigau, (eds.) 2015. *Informe sobre el estado de las Tecnologías del Lenguaje en España*. Accesible en https://www.plantl.gob.es

Soroa, A., G. Rigau, J. Porta, J. Atserias, y X. Gómez Guinovart. 2017. *Plataformas y sistemas de procesamiento lingüístico de alto rendimiento*. Accesible en https://www.plantl.gob.es